Digitizing India's Ancient Texts: AI for Tamil Palm Leaf Manuscript Preservation and Accessibility

NANDHINI PRADEEP, DIVYAR SUBRAMANIAN, and MADHAVAN K. GANAPATHY*, Chidambaram School of Engineering, India

Tamil Palm Leaf manuscripts are invaluable cultural treasures, containing centuries of knowledge in areas such as literature, religion, and medicine. However, their physical fragility presents significant challenges for preservation and access. This study aims to develop a digital archive for Tamil Palm Leaf manuscripts by creating standardized metadata schemas and employing modern digital preservation techniques. The research will explore methods such as high-resolution imaging, AI-based OCR, and machine learning for automated metadata generation to ensure the manuscripts' longevity and facilitate global access. The goal is to provide a centralized, interoperable platform for scholars and the public, enhancing both the preservation and accessibility of Tamil cultural heritage.

Additional Key Words and Phrases: Document Understanding, Tamil Palm leaf Manuscripts, Deep Learning

ACM Reference Format:

Nandhini Pradeep, Divyar Subramanian, and Madhavan K. Ganapathy. 2024. Digitizing India's Ancient Texts: AI for Tamil Palm Leaf Manuscript Preservation and Accessibility. 1, 1 (December 2024), 10 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Tamil Palm Leaf manuscripts are among the oldest and most significant forms of written cultural heritage in South Asia, carrying a wealth of knowledge in various domains such as literature, philosophy, religion, and medicine. These manuscripts, traditionally inscribed on dried palm leaves, have been instrumental in preserving the wisdom of ancient Tamil civilizations for centuries. However, due to their organic nature, Palm Leaf manuscripts are highly susceptible to degradation, including ink fading, physical wear, and environmental damage. As a result, many of these manuscripts remain inaccessible or poorly preserved, posing a significant challenge for researchers, historians, and scholars who seek to study them [2, 6, 24, 27].

In recent years, advancements in digital technologies offer promising solutions for preserving and enhancing access to these invaluable manuscripts. As shown in Fig.1, digital preservation methods, including high-resolution imaging and AI-driven optical character recognition (OCR) [1, 2, 9, 26], can help overcome many of the physical limitations of traditional Palm Leaf manuscripts. However, for such an archive to be truly effective, it requires a robust metadata framework to support both the preservation and the long-term accessibility of these texts [8, 20].

The creation of a digital archive for Tamil Palm Leaf manuscripts, complete with standardized metadata, would facilitate global access to these texts, enabling scholars from diverse fields—such as digital humanities, linguistics, and history—to engage with them in new and innovative ways [3, 11]. This initiative would not only preserve the manuscripts for future generations but also ensure their place within the broader academic and cultural discourse

Authors' Contact Information: Nandhini Pradeep; Divyar Subramanian; Madhavan K. Ganapathy, nandhini.prad1@gmail.com, Chidambaram School of Engineering, Tamil Nadu, India.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

Manuscript submitted to ACM

[5, 13]. This study explores the challenges and opportunities involved in the metadata cataloging and digital archiving of Tamil Palm Leaf manuscripts, focusing on the development of an interoperable system that can make these precious texts accessible to a global audience.

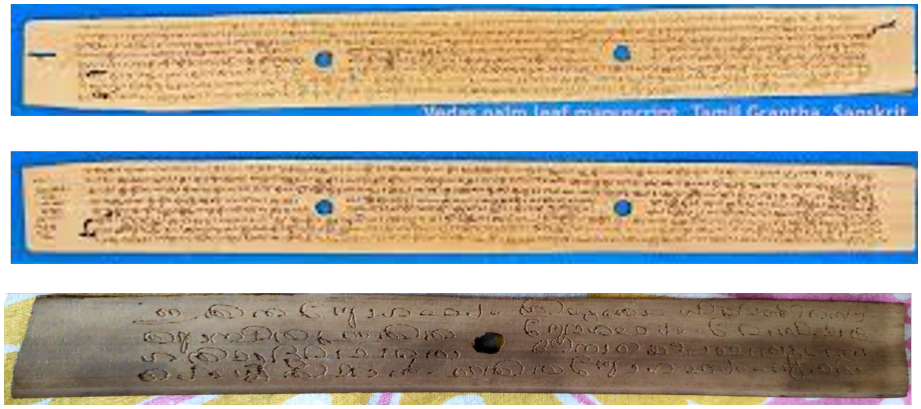

Fig. 1. Illustrative samples of text lines in Palm leaf manuscripts.

2 Literature Review

2.1 Historical Context of Tamil Palm Leaf Manuscripts

Tamil Palm Leaf manuscripts have been a central part of South Indian intellectual and cultural traditions for centuries. These manuscripts, typically created on dried palm leaves, hold a vast range of content, from religious texts like the *Tirukkural* and *Agamas* to medical treatises in *Siddha* medicine, and historical chronicles. The preservation of this knowledge is crucial not only for the understanding of ancient Tamil civilization but also for the broader study of Indian and South Asian cultural history.

Previous studies such have highlighted the unique role of these manuscripts in the transmission of Tamil culture, particularly in oral traditions where manuscripts were the primary medium for codifying and preserving texts [10, 16]. Despite their importance, Tamil Palm Leaf manuscripts have faced significant preservation challenges, including degradation due to climatic conditions, fungal growth, and the physical wear that results from handling. These challenges have led to the loss or inaccessibility of countless manuscripts over time [7, 25].

2.2 Digital Preservation of Manuscripts

With the advent of digital technologies, the field of manuscript preservation has experienced a paradigm shift. Digitization offers a way to safeguard fragile manuscripts while enhancing access for global scholars. A central component of digitization is the creation of high-resolution digital images that capture both the text and intricate material properties of manuscripts, including script style, ink, and even degradation patterns. This digital reproduction ensures that manuscripts can be stored, accessed, and studied without further physical handling, thus preventing further deterioration [19, 23, 25].

Several successful manuscript digitization projects have served as models for digital preservation. Projects like the *Digitization of Tibetan Buddhist Texts* (Tibetan Buddhist Resource Center) and the *Digital Corpus of Sanskrit*

Manuscripts (DCSM) have used high-resolution imaging techniques and integrated them with metadata to create accessible, searchable databases. Similarly, the *Indian Manuscripts Project* by the National Mission for Manuscripts (NMM) has digitized hundreds of ancient Indian manuscripts, providing a critical blueprint for how to approach the digital archiving of culturally significant texts [12, 22].

However, digitizing manuscripts is not without its own set of challenges. Studies indicate that the quality of digitized manuscripts is often affected by the condition of the original material, requiring specialized imaging techniques like infrared or UV scanning to reveal faded or hidden text. Furthermore, while the digitization of manuscripts has been relatively straightforward in cases of printed texts, the complexity of handwritten manuscripts—especially those using non-Latin scripts such as Tamil—poses additional difficulties [18, 21, 30].

2.3 Metadata Standards for Manuscript Cataloging

A fundamental aspect of manuscript digitization is the development of metadata, which is essential for organizing and making the digitized materials searchable. Metadata refers to descriptive information about the manuscript, including title, author, language, date of creation, material properties, and physical condition. In the case of Tamil Palm Leaf manuscripts, additional metadata fields may be required to capture information related to the manuscript's script type, region of origin, and the content's subject matter.

Existing metadata standards for manuscript cataloging include *Dublin Core*, *MARC* (Machine-Readable Cataloging), and *METS* (Metadata Encoding and Transmission Standard). These standards provide a general framework for cataloging manuscripts, but they are often not tailored to the specific needs of Palm Leaf manuscripts. The physical properties of Palm Leaf manuscripts, such as their size, texture, and unique script variations, require adaptations to standard metadata frameworks.

A study on the cataloging of rare manuscripts suggests that while *Dublin Core*'s simplicity and flexibility make it an attractive option, it lacks the specificity needed for describing ancient manuscripts, especially those written in complex scripts. To address this gap, custom metadata schemas, who advocate for incorporating metadata standards that reflect the physical and linguistic characteristics of non-Latin manuscripts [29].

2.4 Case Studies of Digital Archives for Palm Leaf Manuscripts

Several digital archive projects have been initiated to preserve and catalog Tamil Palm Leaf manuscripts, though most are still in early stages. Notable among these is the *Tamil Digital Library* (TDL), which has digitized a substantial portion of Tamil literary and religious manuscripts, including Palm Leaves. The TDL has developed a comprehensive database that combines digitized images with metadata, but challenges in standardizing metadata and ensuring uniform quality of OCR outputs persist.

International collaborations, such as the *South Asian Manuscripts Digital Library* at the University of Chicago, provide additional case studies for how these texts can be preserved and cataloged. The *Sri Lanka National Archives* has also embarked on digitizing Tamil Palm Leaf manuscripts, which involve large-scale efforts to provide multilingual metadata and enhance accessibility via online platforms.

Despite these efforts, gaps remain in metadata consistency and in addressing the unique challenges of digitizing Palm Leaf manuscripts. The work by Chandran (2017) on the *India Heritage Digital Archive* underscores the need for continued refinement in metadata standards and the integration of emerging technologies like machine learning for automated metadata generation.

2.5 The Role of AI in Metadata Generation and Searchability

With the rapid advancement of AI, there are growing opportunities to automate not only the transcription of Tamil Palm Leaf manuscripts but also the generation of relevant metadata. AI technologies, particularly Natural Language Processing (NLP) and machine learning algorithms, can facilitate the extraction of key data points from manuscripts, such as names, dates, and keywords, and associate them with their corresponding metadata [14, 15, 17].

3 Key Challenges and Gaps in the Literature

This section outlines the key challenges and gaps in the existing literature related to the digitization, preservation, and cataloging of Tamil Palm Leaf manuscripts. These challenges encompass issues such as degradation, the limitations of current metadata frameworks, OCR challenges, and the barriers to global access to these important cultural resources.

3.1 Preservation and Degradation of Palm Leaf Manuscripts

Tamil Palm Leaf manuscripts, which are among the oldest forms of written records, are highly susceptible to environmental degradation. Factors like humidity, temperature fluctuations, and exposure to light have led to significant deterioration of the materials over time. Palm leaves, the primary material for these manuscripts, are particularly sensitive to mold, fungal growth, and insect damage. As highlighted by [9], these materials are often brittle, and the ink used to inscribe the manuscripts has faded due to aging, making the text illegible in many cases.

The literature suggests that while traditional preservation methods, such as physical restoration and climate-controlled storage, are used, these methods are not sufficient to protect these manuscripts in the long term. Moreover, they often require specialized expertise and can be costly. The digitization of these manuscripts is considered a vital solution to prevent further physical degradation, but the process itself brings its own challenges. Digitization must be performed with high-resolution imaging techniques that can capture fine details, such as ink traces and wear patterns, to preserve the manuscripts' readability.

A study by [28] emphasizes the need for advanced imaging technologies, such as infrared and UV imaging, to reveal hidden text and improve the quality of digital preservation. However, while these techniques have proven effective in other contexts, they are rarely used for Tamil Palm Leaf manuscripts, and the adoption of such technologies is still in its infancy. Further research is needed to evaluate and refine these technologies for widespread use in Tamil manuscript preservation.

3.2 Metadata and Cataloging Issues

One of the fundamental challenges in digitizing Tamil Palm Leaf manuscripts is the creation of standardized, comprehensive metadata for these unique documents. Metadata for manuscripts typically includes bibliographic information such as title, author, date, language, script, and subject matter. However, due to the diversity of Palm Leaf manuscripts in terms of script variations, language usage, material properties, and regional writing styles, the existing metadata standards such as Dublin Core or MARC (Machine-Readable Cataloging) are often inadequate for cataloging Tamil manuscripts effectively.

For example, the traditional classification of metadata for printed or digitally typed texts does not account for the intricacies of ancient Palm Leaf manuscripts, such as the type of palm leaf used, the method of inscription, and the ink composition. The presence of inscriptions in multiple languages (e.g., Tamil, Sanskrit, and Prakrit) further complicates

the cataloging process. Tamil Palm Leaf manuscripts are often multi-lingual and multi-script, making it difficult to develop a universal framework that accommodates the varied characteristics of these manuscripts.

Moreover, existing digital repositories for Tamil manuscripts often lack consistent metadata structures, making it challenging for scholars to search and retrieve documents based on specific criteria. The absence of standardized metadata formats also limits the ability of institutions to share and collaborate on digitization efforts. A tailored metadata schema that takes into account the regional variations of Tamil script, contextual information related to manuscript contents, and the physical attributes of the manuscripts is critical for improving the accessibility and discoverability of Tamil Palm Leaf manuscripts in digital archives.

3.3 OCR and Text Recognition Challenges

Optical Character Recognition (OCR) is a crucial technology for converting handwritten or printed text into machine-readable formats, which is necessary for digitizing manuscripts. However, the application of OCR to Tamil Palm Leaf manuscripts has proven to be extremely challenging. Several issues contribute to this difficulty, such as the degradation of manuscripts, the variety of handwriting styles, and the complexity of the Tamil script itself.

Tamil Palm Leaf manuscripts often feature handwritten text, which can vary greatly in terms of letterforms, ligatures, and stylizations. The handwriting in these manuscripts can be inconsistent, and the degradation over time further complicates OCR processes. The ink may be faded, smudged, or completely worn away in places, making it difficult for OCR systems to recognize the characters accurately. Additionally, the manuscripts often feature marginalia, annotations, and interlinear text that further complicate the extraction of a clean textual record.

OCR systems designed for modern printed Tamil texts often fail to handle historical manuscripts due to these factors. while some progress has been made in Tamil OCR through machine learning and deep learning approaches, the existing OCR models have limitations when applied to degraded or fragmented manuscripts. Moreover, many current models are not trained on Tamil manuscripts that include other languages or scripts, such as Sanskrit or Prakrit, often seen in Tamil Palm Leaf manuscripts.

The lack of a large, annotated dataset of Tamil Palm Leaf manuscripts for training OCR models is another significant challenge. The absence of such datasets limits the ability of deep learning-based systems to perform well in this domain. There is a clear need for more annotated historical Tamil manuscript datasets, as well as models that can account for the unique handwriting styles and mixed-language content that characterize these manuscripts.

3.4 Data Access and Global Collaboration

The digitization of Tamil Palm Leaf manuscripts faces barriers related to access and global collaboration. Many of these manuscripts are stored in private collections, temples, and local archives, limiting access for scholars and researchers. Moreover, the lack of a unified, global digital platform for Tamil Palm Leaf manuscripts hinders cross-institutional collaboration and prevents researchers from easily discovering and sharing digital copies of these documents.

As [9] discusses, there are several digital archives such as the Tamil datasets and the Indian Manuscripts Project that have made strides in digitizing Tamil manuscripts. However, these archives are often fragmented, with different institutions maintaining separate collections that lack interoperability. The absence of standardized metadata and cataloging practices between these archives complicates the sharing of data and resources. For example, different archives may use different classification systems or metadata schemas, making it difficult for researchers to locate specific manuscripts or compare data across repositories.

There is a growing need for international collaboration in the digitization of Tamil Palm Leaf manuscripts. A unified global digital infrastructure, supported by standardized metadata and open-access platforms, could significantly enhance the visibility and accessibility of these valuable historical documents. Collaborative efforts, such as those initiated by the International Federation of Library Associations and Institutions (IFLA) and the International Council on Archives (ICA), could help address these issues by promoting global standards for manuscript digitization and metadata cataloging.

4 The Role of AI in Tamil Palm Leaf Manuscript Digitization

Artificial Intelligence (AI), particularly in the fields of machine learning (ML) and deep learning (DL), holds significant promise for overcoming many of the challenges associated with Tamil Palm Leaf manuscript digitization. From improving OCR accuracy to automating metadata generation and enhancing image quality, AI can play a crucial role in ensuring the preservation, accessibility, and analysis of these important cultural artifacts.

4.1 AI for OCR and Text Recognition

One of the most immediate applications of AI in Tamil Palm Leaf manuscript digitization is the improvement of OCR for handwritten Tamil text. Traditional OCR models rely on character templates and pattern recognition, which often fail to recognize Tamil characters in historical manuscripts due to factors such as irregular handwriting, ink degradation, and the presence of mixed scripts.

Recent advancements in deep learning have enabled the development of more robust OCR systems, particularly those using Convolutional Neural Networks (CNNs) and Transformer-based models, which can learn complex patterns from large datasets of historical handwritten text. A study by [4] demonstrated that deep learning models, particularly Long Short-Term Memory (LSTM) networks and attention-based models, can achieve better accuracy when applied to Tamil manuscripts compared to traditional OCR systems. These models are capable of recognizing complex letterforms and can adapt to different handwriting styles over time.

To address the problem of mixed-script content (e.g., Tamil and Sanskrit), transfer learning techniques can be employed. Transfer learning allows models trained on other Indian languages, such as Devanagari or Kannada, to be fine-tuned for Tamil. This can significantly improve OCR accuracy, especially when training data for Tamil Palm Leaf manuscripts is limited.

4.2 AI for Metadata Extraction and Cataloging

AI can also be utilized to automate metadata extraction from digitized Tamil Palm Leaf manuscripts. Natural Language Processing (NLP) algorithms, including Named Entity Recognition (NER) and topic modeling, can be employed to extract key metadata from the transcribed text. These include attributes such as the author, manuscript title, place of origin, historical context, and thematic keywords.

The use of AI for metadata extraction can significantly streamline the cataloging process, which is typically time-consuming and labor-intensive. By automating the extraction of key attributes, AI can help create comprehensive and accurate metadata for large collections of Tamil Palm Leaf manuscripts, making them more discoverable and accessible in digital archives.

5 Discussion and Analysis

In this section, we discuss the challenges, opportunities, and implications of using AI for the digitization and preservation of Tamil Palm Leaf manuscripts. The analysis will explore the effectiveness of current technologies, the gaps in the literature, and the potential impact of future innovations.

5.1 Addressing Key Challenges with AI

The literature clearly identifies several pressing challenges in the preservation, cataloging, and recognition of Tamil Palm Leaf manuscripts. These challenges are primarily due to the degradation of materials, the complexity of metadata cataloging, and the limitations of OCR technologies. AI, particularly deep learning and machine learning models, has shown significant potential in addressing these issues.

One of the most promising areas where AI can provide immediate impact is in the improvement of Optical Character Recognition (OCR). The existing OCR models, which often rely on traditional pattern recognition, are inadequate for the degraded, handwritten, and multi-lingual nature of Tamil Palm Leaf manuscripts. As discussed earlier, AI-based models, particularly those based on Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, can better handle the complexities of Tamil script and its variants. These models have the ability to recognize even degraded or partially obscured text, which is a significant improvement over traditional OCR methods.

Furthermore, AI can automate the extraction of metadata from these manuscripts, which is another area where current practices fall short. Traditional cataloging methods require manual input, which is not only time-consuming but also prone to human error. By leveraging Natural Language Processing (NLP) and image recognition techniques, AI can help generate accurate and comprehensive metadata automatically, reducing the manual labor involved in cataloging and ensuring consistency across databases. The automation of metadata extraction also has the added benefit of improving the searchability of manuscripts, which is crucial for global collaboration.

5.2 The Role of AI in Improving Preservation Techniques

Beyond OCR and metadata extraction, AI has the potential to revolutionize the preservation of Tamil Palm Leaf manuscripts. The process of digitization itself often involves the enhancement of images to recover faded text, improve legibility, and even restore damaged portions of the manuscript. AI techniques such as Generative Adversarial Networks (GANs) can be used to reconstruct and enhance the quality of these manuscripts, revealing hidden text and reducing the impact of wear and tear. This aspect of AI-powered restoration could allow us to recover lost or damaged content from manuscripts that would otherwise be unreadable, thus enhancing their preservation for future generations [27].

Additionally, AI could be used to monitor the physical condition of manuscripts over time, by identifying and categorizing types of degradation. This would allow conservators to take preventive measures, such as adjusting environmental conditions or applying targeted restoration techniques.

5.3 Limitations and Challenges in AI Implementation

Despite the potential benefits of AI in Tamil Palm Leaf manuscript digitization, there are several limitations that need to be addressed. A key limitation is the lack of large, high-quality annotated datasets for training AI models. While there have been some efforts to create digitized collections of Tamil manuscripts, these datasets are often small and not sufficiently diverse to train robust deep learning models. The absence of annotated datasets that reflect the full range of

handwriting styles, script variations, and manuscript types poses a significant obstacle to the development of accurate and reliable OCR models.

Another challenge is the need for specialized AI models that can handle multi-lingual and multi-script content. Many Tamil Palm Leaf manuscripts contain inscriptions in multiple languages, including Sanskrit, Prakrit, and even regional dialects, making them more complex to process. Developing AI models that can accurately identify and transcribe these mixed-language manuscripts requires sophisticated techniques that go beyond traditional OCR and involve cross-lingual or transfer learning approaches.

Moreover, the adoption of AI for manuscript digitization also requires substantial computational resources and expertise, which may not be available to all institutions involved in the preservation of Tamil Palm Leaf manuscripts. For many smaller or underfunded institutions, the infrastructure needed to implement AI technologies may be out of reach.

5.4 Opportunities for Future Research

The challenges outlined above point to several important areas for future research. First, there is an urgent need to develop and curate large-scale annotated datasets for Tamil Palm Leaf manuscripts. These datasets should cover a wide range of handwriting styles, scripts, and degradation types to ensure that AI models are trained on a comprehensive and diverse set of examples. Collaborations between universities, cultural institutions, and AI research labs could be instrumental in creating these resources.

Second, AI researchers could explore transfer learning techniques to adapt models trained on other South Asian scripts to Tamil Palm Leaf manuscripts. Given the similarities between Indian scripts, such models could offer a practical solution for improving OCR accuracy with limited annotated data. Furthermore, there is scope for developing multi-modal AI models that can process both the visual and textual components of manuscripts simultaneously. These models could improve the accuracy of metadata extraction by considering both the images of the manuscript and the transcribed text together.

Finally, more research is needed to refine AI-driven image enhancement techniques for the restoration of Tamil Palm Leaf manuscripts. Techniques such as GANs, deep learning-based denoising, and super-resolution imaging could be applied to recover obscured or faded text, improving the legibility and completeness of digitized manuscripts.

6 Conclusion

The digitization and preservation of Tamil Palm Leaf manuscripts represent a critical area of study, both for the preservation of cultural heritage and for advancing our understanding of historical texts. As demonstrated throughout this review, Tamil Palm Leaf manuscripts face a number of significant challenges, including degradation, lack of standardized metadata frameworks, and difficulties in OCR and text recognition. However, the application of Artificial Intelligence (AI) offers promising solutions to many of these challenges.

AI techniques, particularly those based on deep learning, have the potential to significantly improve OCR accuracy for Tamil Palm Leaf manuscripts. By learning from historical handwritten texts, AI models can overcome many of the issues posed by degraded and fragmented manuscripts. Additionally, AI can automate the metadata extraction process, streamlining cataloging efforts and improving the searchability and discoverability of manuscripts in digital archives.

Furthermore, AI has the potential to enhance the preservation of these manuscripts through image enhancement and restoration techniques. GANs and other AI-powered tools can reconstruct faded or damaged text, allowing for the recovery of lost content and improving the overall quality of digitized manuscripts.

Despite these opportunities, several challenges remain, particularly in the creation of large, annotated datasets for training AI models. Collaboration between cultural institutions, AI researchers, and scholars is crucial to address these challenges and ensure that Tamil Palm Leaf manuscripts are digitized, preserved, and made accessible to a global audience.

In conclusion, while AI offers significant promise in overcoming many of the obstacles associated with Tamil Palm Leaf manuscript digitization, continued research and collaboration are essential to fully realize its potential. As AI technologies continue to evolve, they could play a transformative role in ensuring the long-term preservation and accessibility of these invaluable cultural resources.

References

[1] Suganya Athisayamani, A Robert Singh, and A Sivanesh Kumar. 2021. Recurrent neural network-based character recognition system for Tamil palm leaf manuscript using stroke zoning. In *Inventive Communication and Computational Technologies: Proceedings of ICICCT 2020*. Springer, 165–176.

[2] Suganya Athisayamani, A Robert Singh, and T Athithan. 2020. Recognition of ancient Tamil palm leaf vowel characters in historical documents using B-spline curve recognition. *Procedia Computer Science* 171 (2020), 2302–2309.

[3] B Kiran Bala and I Infant Raj. 2020. A novel methodology for identifying the Tamil character recognition from palm leaf. *New Trends in Computational Vision and Bio-inspired Computing: Selected works presented at the ICCVBIC 2018, Coimbatore, India* (2020), 1499–1506.

[4] Thomas M Breuel. 2017. High performance text recognition using a hybrid convolutional-lstm implementation. In *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*, Vol. 1. IEEE, 11–16.

[5] Balaraju Chandramouli and Kokkanti Mallikarjuna. 2020. Enumeration of various Ayurvedic formulations listed in an ancient palm-leaf manuscript of Rayalaseema region in Andhra Pradesh state, India. *Journal of ethnopharmacology* 251 (2020), 112552.

[6] Gayathri Devi S, Subramaniyaswamy Vairavasundaram, Yuvaraja Teekaraman, Ramya Kuppusamy, and Arun Radhakrishnan. 2022. A Deep Learning Approach for Recognizing the Cursive Tamil Characters in Palm Leaf Manuscripts. *Computational Intelligence and Neuroscience* 2022, 1 (2022), 3432330.

[7] Paramasivam Muthan Eswaran, Dinesh Manib, and Sabeenian Royappan Savarimuthu. 2021. RECOGNIZING TAMIL PALM-LEAF MANUSCRIPT CHARACTERS USING HYBRIDIZED HUMAN PERCEPTION BASED FEATURES. *ICTACT Journal on Image & Video Processing* 11, 4 (2021).

[8] J Haritha, VT Balamurugan, KS Vairavel, N Ikram, M Janani, K Indrajith, et al. 2022. CNN based character recognition and classification in tamil palm leaf manuscripts. In *2022 International Conference on Communication, Computing and Internet of Things (IC3IoT)*. IEEE, 1–6.

[9] I Jailingeswari and S Gopinathan. 2024. Tamil handwritten palm leaf manuscript dataset (THPLMD). *Data in Brief* 53 (2024), 110100.

[10] D Udaya Kumar, G Sreekumar, and U a Athvankar. 2009. Traditional writing system in southern India—palm leaf manuscripts. *Design Thoughts* 7 (2009), 2–7.

[11] S Sujith Kumar, B Santhosh, Sm Guruakash, and M Pravin Savaridass. 2023. AI Based Tamil Palm Leaf Character Recognition. In *2023 Third International Conference on Smart Technologies, Communication and Robotics (STCR)*, Vol. 1. IEEE, 1–7.

[12] TR Vijaya Lakshmi, Panyam Narahari Sastry, Ramakrishnan Krishnan, NV Koteswara Rao, and TV Rajinikanth. 2015. Analysis of Telugu palm leaf character recognition using 3D feature. In *2015 International Conference on Computational Intelligence and Networks*. IEEE, 36–41.

[13] TR Vijaya Lakshmi, Panyam Narahari Sastry, and TV Rajinikanth. 2017. A novel 3D approach to recognize Telugu palm leaf text. *Engineering Science and Technology, an International Journal* 20, 1 (2017), 143–150.

[14] Xiaozhong Liu, Chun Guo, and Lin Zhang. 2014. Scholar metadata and knowledge generation with human and artificial intelligence. *Journal of the Association for Information Science and Technology* 65, 6 (2014), 1187–1201.

[15] Bart Magnus, Matthias Priem, Nastasia Vanderperren, Peter Vanden Berghe, Ellen Van Keer, and Rony Vissers. 2025. Metadata creation and enrichment using artificial intelligence at meemoo. *Journal of Digital Media Management* 13, 2 (2025), 110–123.

[16] S Uma Maheswari, P Uma Maheswari, and GR Sai Aakaash. 2024. An intelligent character segmentation system coupled with deep learning based recognition for the digitization of ancient Tamil palm leaf manuscripts. *Heritage Science* 12, 1 (2024), 342.

[17] Gonzalo P Rodrigo, Matt Henderson, Gunther H Weber, Colin Ophus, Katie Antypas, and Lavanya Ramakrishnan. 2018. ScienceSearch: Enabling search through automatic metadata generation. In *2018 IEEE 14th International Conference on e-Science (e-Science)*. IEEE, 93–104.

[18] RS Sabeenian, ME Paramasivam, PM Dinesh, R Adarsh, and Gokul Ravi Kumar. 2017. Classification of handwritten Tamil characters in palm leaf manuscripts using SVM based smart zoning strategies. In *Proceedings of the 2nd International Conference on Biomedical Signal and Image Processing*. 18–21.

[19] RS Sabeeniana, ME Paramasivamb, and PM Dineshc. [n. d.]. Digitalization of Tamil Characters in Palm Leaf Manuscripts Using Image Processing Technique. ([n. d.]).

[20] John G Samuel. 1994. Preservation of palm-leaf manuscripts in Tamil. *IFLA journal* 20, 3 (1994), 294–305.

[21] Panyam Narahari Sastry, Ramakrishnan Krishnan, and Bhagavatula Venkata Sanker Ram. 2010. Classification and identification of Telugu handwritten characters extracted from palm leaves using decision tree approach. *J. Applied Engn. Sci* 5, 3 (2010), 22–32.

[22] M Mohamed Sathik and R Spurgen Ratheash. [n. d.]. Optimal Character Segmentation for Touching Characters in Tamil Language Palm Leaf Manuscripts using Horver Method. ([n. d.]).

[23] R Spurgen Ratheash M Mohamed Sathik. [n. d.]. Line Segmentation Challenges in Tamil Language Palm Leaf Manuscripts. ([n. d.]).

[24] Kavitha Subramani and Murugavalli Subramaniam. 2021. Creation of original Tamil character dataset through segregation of ancient palm leaf manuscripts in medicine. *Expert Systems* 38, 1 (2021), e12538.

[25] Nimol Thuon, Jun Du, and Jianshu Zhang. 2022. Improving isolated glyph classification task for palm leaf manuscripts. In *International Conference on Frontiers in Handwriting Recognition*. Springer, 65–79.

[26] Nimol Thuon, Jun Du, and Jianshu Zhang. 2022. Syllable analysis data augmentation for khmer ancient palm leaf recognition. In *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 1855–1862.

[27] Nimol Thuon, Jun Du, Zhenrong Zhang, Jiefeng Ma, and Pengfei Hu. 2024. Generate, transform, and clean: the role of GANs and transformers in palm leaf manuscript generation and enhancement. *International Journal on Document Analysis and Recognition (IJDAR)* (2024), 1–18.

[28] Athena Van der Perre, Hendrik Hameeuw, Vanessa Boschloos, Luc Delvaux, Marc Proesmans, Bruno Vandermeulen, Luc Van Gool, and Lieve Watteeuw. 2016. Towards a combined use of IR, UV and 3D-imaging for the study of small inscribed and illuminated artefacts. *Lights On... Cultural Heritage and Museums* (2016), 163–192.

[29] EK Vellingiriraj and P Balasubramanie. 2013. Recognition of ancient tamil handwritten characters in palm manuscripts using genetic algorithm. *International Journal of Scientific Engineering and Technology* 2, 5 (2013), 342–346.

[30] TR Vijaya Lakshmi, P Narahari Sastry, and TV Rajinikanth. 2018. Feature selection to recognize text from palm leaf manuscripts. *Signal, Image and Video processing* 12 (2018), 223–229.